Information-Theoretic Distance Measures and a Generalization of Stochastic Resonance

J. W. C. Robinson* and D. E. Asraf†
National Defense Research Establishment FOA, S-172 90 Stockholm, Sweden

A. R. Bulsara‡ and M. E. Inchiosa§
Space and Naval Warfare Systems Center, Code D364, San Diego, California 92152-5001
(Received 26 June 1998)

We show that stochastic resonance (SR)-like phenomena in a nonlinear system can be described in terms of maximization of information-theoretic distance measures between probability distributions of the output variable or, equivalently, via a minimum probability of error in detection. This offers a new and unifying framework for SR-like phenomena in which the "resonance" becomes independent of the specific method used to measure it, the static or dynamic character of the nonlinear device in which it occurs, and the nature of the input signal. Our approach also provides fundamental limits of performance and yields an alternative set of design criteria for optimization of the information processing capabilities of nonlinear devices. [S0031-9007(98)07277-9]

PACS numbers: 05.40.+j, 02.50.Wp, 47.20.Ky, 85.25.Dq

The stochastic resonance (SR) effect [1] is generally quantified in terms of a maximum (as a function of the input noise intensity) of a performance measure, e.g., output signal-to-noise ratio (SNR). However, a complete characterization of system performance, in the presence of underlying randomness, requires knowledge of the whole probabilistic structure of the output, and this can be fully retrieved from the spectral properties only when the system is linear, and operated in the Gaussian noise regime. Hence, any measure of information processing performance of a nonlinear system, based on spectral properties alone, will capture only a few aspects of system performance. Indeed, any definition of SR used as a general measure of information processing performance must utilize the *whole* probabilistic structure of the problem. Moreover, in order to have practical applicability for a given class of problems the definition must represent a fundamental (or universal) limit of performance for this class. For instance, in detection applications a definition of SR must be linked to the limits of detectability of a signal, otherwise the "best" preprocessor (the one giving the best "resonance" according to the SR definition) may not be part of the optimal preprocessor-detector combination.

We consider here the problem of detecting a noise-corrupted signal that has passed through a nonlinear system. An alternative definition for SR-like effects, based on information-theoretic distance measures between probability distributions, is proposed: the minimal achievable probability of error in detection, on the system output. The statistical framework for optimal detection is that of binary hypothesis testing: decide which of two given probability distributions p_0 (the correct distribution under the hypothesis H_0) or p_1 (the correct distribution under the hypothesis H_1) is the correct one for an observed random quantity ξ. In this general formulation, based on *probability distributions*, there is a strong coupling to basic information processing capabilities/limits that can be exploited to characterize system behavior.

Information-theoretic concepts such as entropy and mutual information have been used previously in the study of SR [2]; however, the resonance was quantified via an input-output matching of signals, and not via the separation of output probability distributions. The definition of SR in terms of an optimal detection formulation has been applied recently [3], but only to detectors operating on the spectrum of the signal; this represents a restriction in detector structure. Our results will, therefore, complement and generalize that work in several aspects. The definition used here is completely general and applicable to any type of (static or dynamic) nonlinear device, operated in a noisy environment.

A common (and fundamental) criterion of detector optimality [4] is minimization of the error probability P_E defined as

$$P_E = qP_F + (1 - q)P_M \,, \tag{1}$$

where P_F is the probability of false alarm (decide H_1 when in fact H_0 is true), P_M is the probability of miss (decide H_0 when in fact H_1 is true), and q and $1 - q$ are the *a priori* probabilities of H_0 and H_1, respectively. We see that P_F and P_M together uniquely determine the probability of error P_E. Another common optimality criterion is maximization of the probability of detection

$$P_D = 1 - P_M \tag{2}$$

(decide H_1 when H_1 is true) given a specified maximal false alarm level $P_F \leq \alpha$, which is the Neyman-Pearson (NP) formulation [4]. The optimal detector for both of these criteria (and others) takes the form of a likelihood ratio (LR) test, i.e.,

$$\text{decide } H_0 \text{ if } \frac{p_1(\xi)}{p_0(\xi)} \leq \gamma \text{ vs decide } H_1 \text{ if } \frac{p_1(\xi)}{p_0(\xi)} > \gamma \,, \tag{3}$$

 0031-9007/98/81(14)/2850(4)$15.00

where p_1/p_0 is the likelihood ratio of the two distributions p_0, p_1 that characterize the observed random variable ξ under hypotheses H_0 and H_1, respectively. The threshold γ is chosen as $q/(1-q)$ in the case of minimizing P_E, and is minimized subject to $P_F \leq \alpha$ in the NP case.

Given the LR detector, it is intuitively clear that the best possible detection can be achieved when the probability distributions p_0 and p_1 are separated as much as possible, in some sense; this would correspond to a maximization of the statistical "visibility" of the signal in the noise. Indeed, there is a strong connection between detection, detector performance, and information-theoretic distance measures such as the Ali-Silvey distances [5]. These distance measures are functionals of the likelihood ratio of the form

$$d(p_0, p_1) = h\left[\int f\left(\frac{p_1(\xi)}{p_0(\xi)}\right)p_0(\xi)d\xi\right],$$

where f is a continuous convex function and h is an increasing function. One notable example is obtained for $h(x) = x$ and $f(x) = -\log x$ which yields the relative entropy (or Kullback-Leibler distance). Another one is the $d_{\mathcal{I}}$ divergence defined as

$$d_{\mathcal{I}}(p_0, p_1) = \int \left|(1-q)\frac{p_1(\xi)}{p_0(\xi)} - q\right| p_0(\xi)d\xi \quad (4)$$

for which we have the well-known [5] relation

$$\tilde{P}_E = \tfrac{1}{2} - \tfrac{1}{2}d_{\mathcal{I}}(p_0, p_1), \quad (5)$$

where $\tilde{P}_E$ is the probability of error of the *optimal* detector, i.e., the *minimal probability of error*, which is attained by the LR detector when the threshold is set to $\tilde{\gamma} = q/(1-q)$. Thus, $\tilde{P}_E$ and $d_{\mathcal{I}}$ uniquely determine each other via a monotone function and are equivalent. This establishes the connection between information-theoretic distances and detection. Moreover, it follows that maximization of $d_{\mathcal{I}}$ in (4) over parameters for the output of a device is equivalent to a minimization of $\tilde{P}_E$ for detection on the output, and these two quantities both represent relevant theoretical limits of performance for a device used as a preprocessor in detection.

It is a standard property of Ali-Silvey distances that a nonlinear transformation cannot increase the value of any given distance (e.g., $d_{\mathcal{I}}$) between two distributions [5]. Thus, it is clear that a possible alternative definition of SR is the local maximization (over parameters) of $d_{\mathcal{I}}$ (or minimization of $\tilde{P}_E$) for the output distributions corresponding to H_0 and H_1, respectively, given a fixed value of $d_{\mathcal{I}}$ (or $\tilde{P}_E$) for the corresponding input distributions. Alternatively, an output-input ratio (which cannot exceed unity) of $d_{\mathcal{I}}$ divergences could be considered.

In the NP formulation, detector performance is often evaluated via plots of the so-called receiver operating characteristics (ROCs) [4] in which P_D is expressed as a function of P_F. To generate a ROC, one lets the threshold γ in (3) run through all values in $(0, \infty)$; this yields all possible pairs of P_F and P_D. In particular, for the ("optimal") threshold $\tilde{\gamma} = q/(1-q)$ the resulting P_E $(= \tilde{P}_E)$ is directly linked to the $d_{\mathcal{I}}$ divergence via (1), (2), and (5) (since (1) and (2) then deliver optimal values). This means that from a family of ROCs indexed by some parameter (such as input noise variance), one can obtain a plot of how $d_{\mathcal{I}}$ varies by simply picking from each ROC the value of P_E obtained for the threshold $\tilde{\gamma}$ and plotting the corresponding $d_{\mathcal{I}}$ against the parameter. Thus, the ROCs for the optimal detector contain all the information needed to determine an information-theoretic limit for separation between signal and noise.

We now compute the ROCs for a specific nonlinear device, the single junction (rf) SQUID operating in the dispersive (i.e., *non*hysteretic) mode [6,7]. The magnetic flux $x(t)$ (expressed as the dimensionless ratio of the actual magnetic flux to the flux quantum $\Phi_0 \equiv h/2e$) through the loop can be described by the equation of motion $\tau_L \frac{dx}{dt} = -U'(x) + x_e$, where $U(x) = \frac{1}{2}x^2 - \frac{\beta}{4\pi^2}\cos(2\pi x)$ is the potential energy function and $\beta \equiv 2\pi L I_c/\Phi_0$ is the nonlinearity parameter (L and I_c are the loop inductance and the junction critical current, respectively). The externally applied magnetic flux component $x_e(t) = x_0 + x_i(t) + y(t)$ is the sum of a dc level $x_0 \equiv \frac{1}{2}$ (to obtain a symmetric transfer characteristic), an input signal $x_i(t)$ (to be specified later), and noise $y(t)$. The noise, regardless of its origin, is usually effectively band limited by the SQUID bandwidth τ_L^{-1}; it is modeled as zero-mean Gaussian, with an exponentially decaying normalized correlation coefficient $R(\tau) = \sigma^{-2}\langle y(t)y(t+\tau)\rangle_t = e^{-|\tau|/\tau_c}$, τ_c being the noise correlation time, and σ the standard deviation. For our results (and in many practical applications) the noise bandwidth τ_c^{-1} is considerably larger than the signal bandwidth, so that the noise $y(t)$ appears white relative to $x_i(t)$. In most practical cases we also have $\tau_c^{-1} \ll \tau_L^{-1}$, so that the equation of motion reduces to the quasistatic form (considered through the remainder of this work) $U'(x) = x_e$.

The SQUID output is characterized by the "shielding flux" $x_s(t) \equiv x(t) - x_e(t)$. From the quasistatic equation of motion we can obtain the input-output transfer characteristic $x_s(t) = g(x_i(t))$ by solving for x_s as a function of x_i [with $x_0 = \frac{1}{2}$ and $y(t) = 0$]. This has been done analytically [6,7] in the nonhysteretic regime $0 \leq \beta < 1$, to which we confine ourselves. The transfer function g is plotted in Fig. 1; g is periodic in x_i (only one cycle shown), the slope of the central "linear" regime near the origin increases, and the minimal distance Δx_i between extreme points decreases, respectively, with increasing β.

In our model, we have analytically (using the formulae for transformation of probability densities through a nonlinearity) calculated the ROCs for the NP optimal detector based on the SQUID output for different values of the parameters of the input noise, signal, and the nonlinearity. Here, the measured quantity ξ is the output $x_s(t)$ at a fixed time t when the input $x_i(t)$ has a fixed

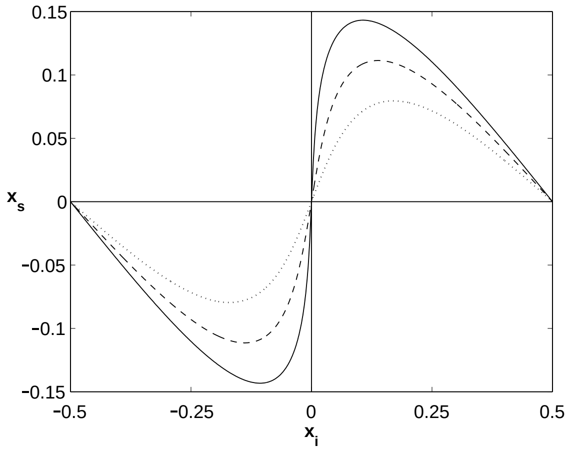

FIG. 1. One period of the rf SQUID transfer characteristic $x_s = g(x_i)$, with $x_0 = \frac{1}{2}$ [and $y(t) = 0$], for $\beta = 0.5$ (dotted line), 0.7 (dashed line), and 0.9 (solid line). The minimum and maximum are separated by a (β dependent) distance Δx_i.

value equal to 0 under H_0 (no signal) and μ (>0) under H_1 (signal), respectively. Thus, under H_0 the output is characterized by a probability distribution $p_0^{(\mathrm{o})}$ and under H_1 by another distribution $p_1^{(\mathrm{o})}$. The corresponding distributions $p_0^{(\mathrm{i})}, p_1^{(\mathrm{i})}$ on the noise-corrupted input $x_i(t) + y(t)$ under H_0 and H_1, respectively, are two Gaussian distributions with the same standard deviation σ but with differing means 0 and μ, respectively. We have varied the input noise variance σ^2 but changed μ accordingly so that on the noise corrupted input $\tilde{P}_E$, $d_{\mathcal{I}}$, and SNR (here defined as μ^2/σ^2) have all been held constant in each family of ROCs. (This is possible for a Gaussian distribution.) The results are displayed in Figs. 2 and 3.

Two asymptotes exist in all the ROC families. One is obtained when the input noise variance approaches zero and the ROCs tend to those for two Gaussian distributions (since the transfer function then acts essentially linearly), as can be seen in the leftmost part of all ROC families. The other asymptote is obtained when the noise variance tends to infinity and $p_0^{(\mathrm{o})}, p_1^{(\mathrm{o})}$ both collapse to the distribution obtained by transforming a uniform distribution through one period of the nonlinearity. In this case the ROCs become a straight line with unit slope as in the rightmost part of all ROC families. From the ROCs the $d_{\mathcal{I}}$ divergence is obtained via (1), (2), and (5) by reading off the P_F, P_D pairs along a curve (corresponding to the optimal threshold $\tilde{\gamma}$) on the ROC surface, as indicated above for a one-parameter family of ROCs.

In Fig. 3 we see clear evidence of "resonant" behavior in terms of local maximization of output $d_{\mathcal{I}}$ for all values of β greater than 0.7, with a global maximum for input variance zero. In the zero variance limit, the value of the output $d_{\mathcal{I}}$ obtained is the same as the input value (except possibly for the highest β's). This is to be expected since the input distributions $p_0^{(\mathrm{i})}, p_1^{(\mathrm{i})}$ are highly localized when σ^2 (and thereby μ) is small, so that the linear

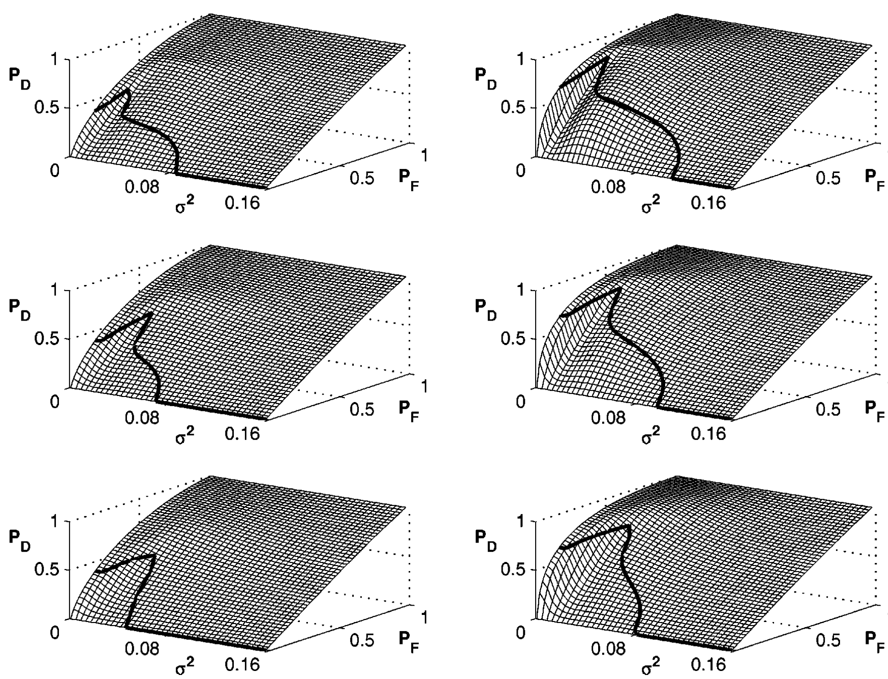

FIG. 2. ROCs for the optimal (LR) detector on the output of an rf SQUID with Gaussian noise on the input and different input $d_{\mathcal{I}}$-SNR levels. The *a priori* probability q for noise only is always 0.6, and the corresponding value $1 - q$ for signal plus noise is 0.4. The left column of ROCs is for constant input $d_{\mathcal{I}} = 0.318$ (SNR = 0.5) and the right column is for input $d_{\mathcal{I}} = 0.538$ (SNR = 2), where β in each column is 0.9 (top), 0.7 (middle), and 0.5 (bottom), respectively. For each σ^2, the $d_{\mathcal{I}}$ divergence is obtained from the relations (1), (2), and (5) by reading off the P_F, P_D pairs obtained from the curve (dark solid lines) on the surface that correspond to the optimal threshold $\tilde{\gamma} = q/(1 - q)$.

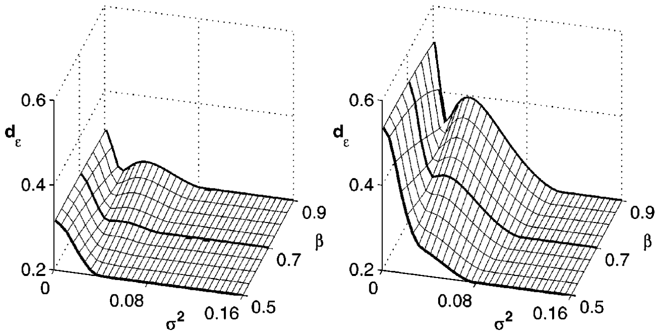

FIG. 3. Output $d_{\mathcal{E}}$ divergence vs input variance σ^2 and β for two different levels of input $d_{\mathcal{E}}$: 0.318 (left) and 0.538 (right) (SNR = 0.5, 2.0, respectively). The curves corresponding to the ROCs in Fig. 2 are marked (dark solid line).

action of the transfer function g near the origin dominates, and linear transformation preserves $d_{\mathcal{E}}$ divergence. In the large noise limit $p_0^{(o)}, p_1^{(o)}$ become identical and the output $d_{\mathcal{E}}$ assumes its minimal value $|1 - 2q|$. The local maxima in the $d_{\mathcal{E}}$ curves occur when $p_0^{(i)}, p_1^{(i)}$ obtain such a scale and position that the nonlinearity redistributes the probability mass to the output most efficiently, in the sense of separating $p_0^{(o)}, p_1^{(o)}$. This matching between $p_0^{(i)}, p_1^{(i)}$ and g depends, significantly, on local properties (e.g., slope and curvature) of the latter, in different regions.

The plots reveal that the local maximum in output $d_{\mathcal{E}}$ occurs when the mean (and mode) μ of the input distribution $p_1^{(i)}$ lies slightly to the right of $\Delta x_i/2$, and the value of μ at resonance moves to the *right* as the input $d_{\mathcal{E}}$ increases ($p_0^{(i)}, p_1^{(i)}$ become more localized). This is related to the fact that the slope of g to the right of $\Delta x_i/2$ is less steep than to the left which gives a greater concentration effect when transforming probability mass. For increasing input $d_{\mathcal{E}}$ the maximum in output $d_{\mathcal{E}}$ becomes more pronounced since more localized (given μ) input distributions can better match local properties of g. It also becomes more pronounced for higher β's, mainly because the first maximum in g is then higher, yielding a greater range on the output and thus greater possibilities for redistributing probability mass. The local minima in the output $d_{\mathcal{E}}$ curves occur roughly when μ is at $\Delta x_i/2$. When β becomes too small no local maximum exists, essentially because the height of the first maximum in g decreases rapidly with β and thereby compresses the output distribution $p_1^{(o)}$ to a region where most of $p_0^{(o)}$ resides. The behavior of P_D as a function of noise strength σ^2 for constant P_F in the ROCs in Fig. 2 is qualitatively similar to the $d_{\mathcal{E}}$ behavior of Fig. 3 and also shows several qualitative similarities with earlier results [3] obtained for a very different system and detector. The resonance behavior displayed here is reminiscent of our earlier observations [7] on the SNR response of the same system under periodic forcing. We conjecture that several parts of this behavior are "generic" for nonlinear systems; in particular, we expect to see it in hysteretic devices.

In conclusion, we have demonstrated that SR-like effects are present even in the most basic instances of information processing in a nonlinear device, and we have related these effects to fundamental limits of performance in detection. We used the $d_{\mathcal{E}}$-divergence curves derived from ROC curves for the optimal detector operating on the output of a nonhysteretic SQUID to demonstrate "resonant" behavior and found stronger "resonances" for higher degrees of nonlinearity. In the small and large noise limits, respectively, we saw the expected asymptotes, and in all cases the output distance $d_{\mathcal{E}}$ was found to be maximal in the zero noise limit and minimal in the large noise limit, as predicted by the properties of the $d_{\mathcal{E}}$ divergence. The results were derived for a 1D case; however, the ideas are quite general. In fact, the dimension of the underlying detection problem (in the sense of the number of samples of the observed output) is not significant, and the ideas are (using more elaborate theory/computational methods) applicable also to infinite-dimensional cases with continuous time observations and more complex signals. This will be the subject of future publications.

J. W. C. R. and D. E. A. acknowledge support from FOA, Project No. E6022 Nonlinear Dynamics; A. R. B. and M. E. I. acknowledge support from the Office of Naval Research.

*Electronic address: john@sto.foa.se
†Electronic address: daniela@sto.foa.se
‡Electronic address: bulsara@spawar.navy.mil
§Electronic address: inchiosa@spawar.navy.mil

[1] For good overviews, see K. Wiesenfeld and F. Moss, Nature (London) **373**, 33 (1995); A. Bulsara and L. Gammaitoni, Phys. Today **49**, 39 (1996); L. Gammaitoni, P. Hanggi, P. Jung, and F. Marchesoni, Rev. Mod. Phys. **70**, 1 (1998).

[2] M. Stemmler, Network **7**, 687 (1996); A. Bulsara and A. Zador, Phys. Rev. E **54**, R2185 (1996); C. Heneghan, C. Chow, J. Collins, T. Imhoff, S. Lowen, and M. Teich, Phys. Rev. E **54**, R2228 (1996); A. Nieman, B. Shulgin, V. Anishchenko, W. Ebeling, L. Schimansky-Geier, and J. Freund, Phys. Rev. Lett. **76**, 4299 (1996); F. Chapeau-Blondeau, Phys. Rev. E **55**, 2016 (1997).

[3] M. Inchiosa and A. Bulsara, Phys. Rev. E **53**, R2021 (1996); M. Inchiosa, A. Bulsara, J. Lindner, B. Meadows, and W. Ditto, in *Chaotic, Fractal, and Nonlinear Signal Processing,* edited by R. A. Katz, AIP Conf. Proc. No. 375 (AIP, New York, 1996); M. Inchiosa and A. Bulsara, Phys. Rev. E **58**, 115 (1998); V. Galdi, V. Pierro, and I. Pinto, Phys. Rev. E **57**, 6470 (1998).

[4] See, e.g., H. van Trees, *Detection, Estimation, and Modulation Theory* (Wiley, New York, 1978).

[5] S. Ali and D. Silvey, J. R. Stat. Soc., Ser. B **28**, 131 (1966); G. Orsak and B. Paris, IEEE Trans. Inf. Theor. **41**, 188 (1995).

[6] A. Barone and G. Paterno, *Physics and Applications of the Josephson Effect* (Wiley, New York, 1982).

[7] M. Inchiosa, A. Bulsara, A. Hibbs, and B. Whitecotton, Phys. Rev. Lett. **80**, 1381 (1998).